EXHIBIT 2

UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

DAVID FLOYD, *et al.*, Plaintiffs, -against- CITY OF NEW YORK, Defendant.	08 Civ. 1034 (AT)
KELTON DAVIS, *et al.*, Plaintiffs, -against- CITY OF NEW YORK, *et al.*, Defendants.	10 Civ. 0699 (AT)

DECLARATION OF JACK GLASER
IN SUPPORT OF PLAINTIFFS' LETTER TO THE COURT

I, Jack Glaser, pursuant to 28 U.S.C. § 1746 and subject to penalties of perjury, state the following is true and correct:

1. I have been a Professor at the Richard & Rhoda Goldman School of Public Policy at the University of California, Berkeley since 2000, where I teach graduate-level courses in quantitative methods and advanced policy analysis, as well as electives on psychological bases and policy implications of prejudice and discrimination. I am one of the principal investigators on a National Science Foundation- and Google-funded project to build a National Justice Database of police stops and use of force incidents. I serve as a consultant to California's Department of Justice, advising on the analysis and interpretation of statewide police stop data that is collected under the Racial and Identity Profiling Act (AB953). I also serve as an advisor to the Office of the Governor of California to develop statewide reforms to police use of force policies.

2. I have been retained by the *Floyd* Plaintiffs as a testifying expert.

3. As part of my work as an expert in this case, I have reviewed the Fourteenth Report of the Independent Monitor on NYPD Social Distancing Enforcement in 2020 (*Floyd* ECF No. 863-1), filed on October 12, 2021 ("Report"), as well as previous drafts of the Report dated June 28, 2021 and September 9, 2021.

4. I submit this declaration in support of Plaintiffs' objections to the Report.

5. I have reviewed the Declaration of Jeffrey Fagan ("Fagan Decl.") and I agree with it. I write separately to add some additional points.

Conceptual Problems

6. One fundamental problem in the Report is its misuse of the term "enforcement." By differentiating between "interactions" and "enforcement," the Monitor's report gives the impression that encounters that do not result in a summons or arrest are not "enforcement." But "enforcement" is a much broader construct than just sanctioning. Certainly, encounters that result in a warning represent enforcement (but they are not included in the Report's analyses). Arguably, any surveillance and detention—even consensual—is enforcement. This lack of analysis of what are probably the most frequent enforcement actions—and the highest in discretion—renders the conclusions incomplete, at best. Worse, they are "selecting on the dependent variable" by only looking at encounters that result in sanctions. Those encounters that result in sanctions are least likely to exhibit racial disparities because the offending behavior largely determines the officers' behavior— giving a summons or making an arrest. The more ambiguous cases, wherein there is less or no evidence to support a summons or arrest, are those that tend to be more influenced by racial or ethnic bias. This plays out in the findings from NYPD stop, question, and frisk ("SQF") data from past years showing that search yield rates tend to be higher for Whites who were

searched than for Blacks or Latinos who were searched.[1] By only looking at incidents that resulted in sanctions, no inference can be made about disparities in the many incidents that did not lead to sanctions.

7. As Jeffrey Fagan's Declaration points out, there are serious problems with the use of social distancing Calls for Service ("CFS") as a benchmark for social distancing violations. *See* Fagan Decl. at ¶¶ 4–7. We have no reason to think CFS is a reasonable proxy for actual or observable violations, given established patterns of racially disparate proactive policing, and given the low likelihood that CFS accurately represent the racial distribution of officer-initiated interactions. *See id.* at ¶ 5. There is good reason to believe that much social distancing enforcement is done on observation by officers. Until the Monitor's team can make a compelling case that CFS are a very strong predictor of rates of all encounters—including officer-initiated encounters—any conclusions about disparity based on comparisons to CFS rates are not compelling.

8. If CFS were a satisfactory benchmark for the racial distribution of all encounters, it would not be a problem that the Report lacks data about when social distancing-related interactions do *not* lead to enforcement actions. However, because CFS cannot be assumed to be an accurate benchmark, the lack of this information means that it is not possible to compellingly estimate rates of enforcement. Encounters without enforcement actions are likely to be many and have potential for racial disparity.

[1] See e.g., Amanda Charbonneau & Jack Glaser, *Suspicion and Discretion in Policing: How Laws and Policies Contribute to Inequity*, 11 UC IRVINE LAW R. 1327 (2021).

Statistical Problems

9. The Report's analysis of bias violates a basic tenet of statistics by "affirming the null hypothesis"—potentially committing a false negative error with respect to the actual rate of summonses relative to the rate predicted based on CFS. Comparing precincts sorted into deciles as a function of the non-White percentage of the precinct's resident population, their data show, descriptively, that the most non-White precincts have rates of summonses that exceed what would be expected based on the number of social distancing CFS, while the least non-White precincts have rates of summonses that are below what would be expected based on CFS. However, because their analysis does not find these differences to reach conventional thresholds for "statistical significance," the Report concludes that there is no evidence of racially disparate enforcement. One problem with this conclusion is that, with roughly seven or eight precincts in each decile, these comparisons have low statistical power, i.e., they are prone to false negative error— failing to detect a difference when there is one. The analysts seek to overcome this limitation by generating more robust estimates with random shuffles of the data. Nevertheless, the Report concludes that there are no statistically significant differences. What is problematic is that statistical significance testing (a.k.a. Null Hypothesis Significance Testing) is designed to allow an affirmative claim when a difference is found (when the "null hypothesis" is rejected), but it is not equipped, nor should it be employed, to affirm the null hypothesis when it cannot be rejected. In other words, a statistically nonsignificant result, i.e., with a p value greater than .05, is not evidence of equality.

10. To make matters worse, in several key comparisons in the most White and non-White precinct deciles, the estimated probability that the difference (between rates of summonses predicted by rates of CFS and actual rates of summonses) is due to chance is quite low. As the

Report states, with respect to the difference observed for the most non-White decile of precincts (the seven most non-White precincts in the City), "The difference of 6.37 extra summonses is not statistically significant according to permutation inference (shuffling social distancing summonses at random 1,000 times to different combinations of precincts). An absolute difference equal to or greater than 6.37 would occur by chance 71 out of 1,000 times, or 7 percent of the time." *See* Report at 17. In other words, this disparity would rarely occur by chance. The essence of significance testing is making an inference about the likelihood that a difference is real or appeared by chance, perhaps due to unrepresentative sampling. The usual ("$p < .05$") standard is that a difference in a sample would have to occur by chance less than five percent of the time in order to consider it "statistically significant" – meaning very unlikely to be a false positive. Here, what is effectively a $p = .07$ is taken as the absence of evidence of a difference. This seems more likely to be a false negative than a false positive. In fact, with respect to summonses (see Figure 1 and Table 4 of the Report), most of the estimates of the likelihood that the observed difference would have occurred by chance are relatively low, including 0.026 for the third most White decile of precincts, where actual summonses were substantially lower than predicted (based on CFS). The general pattern that can be observed in Table 4 and Figure 1 is that, for the most White deciles, the rate of actual summonses tends to be below what would be predicted by the rate of CFS, and for the most non-White deciles, actual summonses tend to exceed predicted summonses. This is a pattern that strongly suggests racially disparate enforcement, even by the analysts' own chosen method of benchmarking with CFS. And yet the conclusion appears to be negative.

11. The burden of proof in significance testing is on finding a difference. The failure to find a difference is not a finding of equality. To further complicate matters, given that racial disparities in enforcement have been found in larger NYPD data sets, it seems reasonable for a

finding of disparity to not hold a particularly high burden of proof. The problem is further compounded by the fact that the data are population-level data — these are all of the incidents known to the Department. Significance testing serves the purpose of making an inference about a population based on a sample. It requires that the sample is random and that it is smaller than the population, and the significance level (the "p-value") is an estimate of the probability that one is committing a false positive error (concluding there is a difference when there isn't one), if one infers from the sample that there is a difference between groups in the population. For population-level data, any observed difference cannot be a false positive. The question should be about *substantive* significance (how big and influential is the difference?), rather than *statistical* significance. The Report compounds these two problems, inferring equivalence based on non-significant difference, and misapplying the statistical significance standard to population-level data.

12. A further problem rests in the Report's dismissive discussion of a robust and statistically significant simple correlation between precinct percent non-White and "probability of enforcement"— the ratio of enforcement actions to CFS. *See* Report at Table 3 and surrounding text. They find that there are more enforcement actions per CFS the more non-White a precinct is ($r = 0.316$; $p = 0.005$). Legitimate concerns are raised because some of the very non-white precincts had single incidents that involved large numbers of summonses or arrests, potentially skewing the correlation (an outlier effect). This is effectively addressed, however, by an analysis that looks at the correlation between these variables in terms of their rank ordering (instead of raw scores). The effect of this is to mitigate the outliers so that a very large number of enforcement actions will not skew the correlation. Nevertheless, the correlation between the ranks is even larger ($r = 0.479$; $p < 0.001$). A 0.5 correlation is considered "large" by scientific standards, and

this one is highly statistically significant. We need not be concerned about "reverse causality" – it is highly improbable that rates of enforcement caused precincts to be more or less non-White. There could be other variables that explain this correlation, and the Report states several possibilities, but none are particularly compelling. They then present the analyses (described above) carving up precincts into non-White deciles to try to address this, but it is not clear how this approach is advantageous. In both the simple correlation and the decile approach, they are benchmarking on CFS. The latter approach is a multivariate one, controlling for other potential causal mechanisms, but a clear case for why the effects are less significant, aside from the loss of statistical power resulting from making comparisons within deciles of small numbers of precincts, is not made.

13. In sum, I find the statistical tests for racial disparities in social distancing enforcement to be problematic. They are too ready to dismiss the import of a large simple correlation between percent non-White and excessive enforcement; violate statistical principles by affirming the null hypothesis; and compound this error by conducting significance testing on population level data when it is intended for making inferences about populations based on sample data. Where the data suggest racially disparate enforcement, the Report seems to stretch to dismiss this evidence and embrace nonsignificance as proof of equivalence.

Dated: December 1, 2021
San Francisco, California

Jack Glaser